METHOD AND ARRANGEMENT FOR DETERMINING SPECTRAL SPEECH CHARACTERISTICS IN A SPOKEN EXPRESSION

The invention is directed to a method and to an arrangement for determining spectral speech characteristics in a spoken expression.

In a concatenative speech synthesis, individual sounds are combined from speech data banks. In order to thereby obtain a speech curve that sounds natural to the human ear, discontinuities must be avoided at the points were the sounds are combined (concatenation points). In particular, the sounds are thereby phonemes of a language or a combination of a plurality of phonemes.

[1] discloses a wavelet transformation. In wavelet transformation, a wavelet filter assures that a respective high-pass part and low-pass part of a following transformation stage completely restore a signal of a current transformation stage. A reduction of the resolution of the high-pass part or, respectively, low-pass part thereby ensues from one transformation stage to the next (English art term: "sub-sampling"). In particular, the plurality of transformation stages is finite due to the sub-sampling.

US-A-5528725 discloses a method for speech recognition with wavelet transformations.

EP-A-0519802 discloses a method for speech synthesis that adapts speaker-specific characteristics in view of a natural sounding concatenation of speech sounds.

The object of the invention is comprised in specifying a method and an arrangement for determining spectral speech characteristics with whose assistance, in particular, a speech output that sounds natural can be determined.

This object is achieved according to the features of the independent claims.

A method for determining spectral speech characteristics in a spoken expression is recited in the scope of the invention. To that end, the spoken expression is digitalized and subjected to a wavelet transformation. The speaker-specific characteristics are determined on the basis of different transformation stages of the wavelet transformation.

One advantage, in particular, is thereby that the expression is divided in the wavelet transformation with a high-pass filter and a low-pass filter and different high-pass parts or, respectively, low-pass parts of different transformation stages contain speaker-specific characteristics.

The individual high-pass parts or, respectively, low-pass parts of different transformation stages stand for predetermined speaker-specific characteristics, whereby both high-pass part as well as low-pass part of a respective transformation stage, i.e. the respective characteristic, can be modified separately from other characteristics. When, in inverse wavelet transformation, the original signal is in turn combined from the respective high-pass and low-pass parts of the individual transformation stages, then it is assured that it is exactly the desired characteristic that has been modified. It is thus possible to modify certain predetermined peculiarities of the expression without the rest of the expression being thereby influenced.

One development is comprised therein that the expression is windowed before the wavelet transformation, i.e. a predetermined set of samples are cut out, and is transformed into the frequency domain. In particular, a fast-Fourier-Transformation (FFT) is employed for this purpose.

A further development is comprised therein that a high-pass part of a transformation stage is split into a real part and an imaginary part. The high-pass part of the wavelet transformation corresponds to the difference signal between the current low-pass part and the low-pass part of the preceding transformation stage.

In particular, one development is comprised therein that the number of transformation stages of the wavelet transformation to be implemented be defined in that a constant part of the expression is contained in the last transformation stage, which is composed of series-connected low-passes. The signal as a whole can then be presented by its wavelet coefficients. This corresponds to the complete transformation of the information of the signal excerpt into the wavelet space.

When, in particular, only the respective low-pass part is further-transformed (with a high-pass and a low-pass filter), then the difference signal remains as high-pass part of a transformation stage, as explained above. When difference signals (high-pass parts) are accumulated over the transformation stages, then the information of the spoken expression without constant part is obtained in the last transformation stage as cumulative high-pass part.

In the scope of an additional development, the speaker-specific characteristics can be identified as:

a) Basic Frequency:

The oscillation of the high-pass part of the first or of the second transformation stage of the wavelet transformation allows the basic frequency of the expression to be recognized. The basic frequency indicates whether the speaker as a man or a woman.

b) Shape of the Spectral Envelope:

The spectral envelope contains information about a transfer function of the vocal tract in the articulation. The spectral envelope is dominated by the formants in a voiced region. The high-pass part of a higher transformation stage of the wavelet transformation contains this spectral envelope.

c) Spectral Tilt (Huskiness):

The huskiness in a voice is visible as negative slope in the curve of the penultimate low-pass part.

The speaker-specific characteristics a0 through c) are of great significance in the speech synthesis. As initially mentioned, large sets of actually spoken expressions from which exemplary sounds are excerpted and later combined to form a new word are used in concatenative speech synthesis (synthetic speech). Discontinuities between combined sounds are thereby disadvantageous since the human ears perceives these as being unnatural. In order to oppose discontinuities, it is advantageous to directly acquire the perceptively relevant quantities and, potentially, to compare and/or adapt them to one another.

This can occur by direct manipulation in that a speech sound is adapted at least in terms of its speaker-specific characteristics, so that it is not perceived as being disturbing in the acoustic context of the concatenatively linked sounds. It is also possible to direct the selection of a suitable sound such that speaker-specific characteristics of sounds to be linked match one another as well as possible, for example that the same or similar huskiness is inherent in the sounds.

One advantage of the invention is comprised therein that the spectral envelope reflects the articulation tract of the speaker and is not supported on formants like, for example, a pole-point model. Further, no data are lost as non-parametric representation in the wavelet transformation, the expression can always be completely reconstructed. The data proceeding from the individual transformation stages of the wavelet transformation are linearly independent of one another, can thus be influenced separately from one another and be recombined later to form the influenced expression -- loss-free.

Further, an arrangement for determining spectral speech characteristics is recited that comprises a processor unit that is configured such that an expression can be digitalized. Subsequently, the expression is subjected to a wavelet transformation and speaker-specific characteristics are determined on the basis of different transformation stages.

This arrangement is particularly suited for the implementation of the nc method or one of its developments explained above.

Developments of the invention also derive from the dependent claims.

Exemplary embodiments of the invention are presented and explained below on the basis of the drawing.

Shown are:

Figure 1 a wavelet function;

Figure 2 a wavelet function subdivided according to real part and imaginary part;

Figure 3 a cascaded filter structure that represents the transformation steps of the wavelet transformation;

Figure 4 low-pass parts and high-pass parts of different transformation stages;

Figure 5 steps of the concatenative speech synthesis.

Figure 1 shows a wavelet function that is defined by

$$\psi(f) = c \cdot \left(1 - \left(\frac{f}{\sigma}\right)^2\right) \cdot e^{-\frac{1}{2}\cdot\left(\frac{f}{\sigma}\right)^2} \qquad (1),$$

whereby

f references the frequency,

σ references a standard deviation, and

c references a predetermined norming constant.

In particular, the standard deviation σ is defined by the prescribable location of the sideband minimum 101 in Figure 1.

Figure 2 shows a wavelet function with a real part according to Equation (1) and a Hilbert transform H of the real part as imaginary part. The complex wavelet function thus derives as

$$\Psi(f) = \psi(f) + j \cdot H\{\psi(f)\} \qquad (2).$$

The constant c from Equation (1) is employed in order to norm the complex wavelet function:

$$\int_{-\infty}^{\infty} \Psi(f) \cdot \overline{\Psi}(f)\, df = 1 \qquad (3),$$

whereby $\overline{\Psi}$ references the conjugated-complex wavelet function.

Figure 3 shows the cascaded application of the wavelet transformation. A signal 301 is filtered both by a high-pass HP1 302 as well as by a low-pass TP1 305. In particular, a sub-sampling thereby occurs, i.e. the plurality of values to be stored is reduced per filter. An inverse wavelet transformation assures that the original signal

301 can in turn be reconstructed from the low-pass part TP1 305 and the high-pass part HP1 304.

Filtering in the high-pass HP1 302 is separated according to real part Re1 303 and imaginary part Im 1 304.

Following the low-pass filter TP1 305, the signal 310 is filtered anew both by a high-pass HP2 306 as well as by a low-pass TP2 309. The high-pass HP2 306 again comprises a real part Re2 307 and an imaginary part Im2 308. Following the send transformation stage 311, the signal is filtered again, etc.

When a (FFT-transformed) short-time spectrum with 256 values is assumed, then eight transformation steps are implemented (sub-sampling rate: 1/2) until the signal from the last low-pass filter TP8 corresponds to the constant part.

Figure 4 shows various transformation stages of the wavelet transformation, divided according to low-pass parts (Figures 4A, 4C and 4E) and high-pass parts (Figures 4B, 4D and 4F).

The basic frequency of the spoken expression can be seen from the high-pass part according to Figure 4B. In addition to the fluctuations in the amplitude, a dominating periodicity in the wavelet-filtered spectrum, the basic frequency of the speaker, can be clearly recognized. On the basis of the basic frequency, it is possible to adapt predetermined expressions to one another in the speech synthesis or to define suitable expressions from a data bank with predetermined expressions.

The formants of the voice signal excerpt (the length of the voice signal excerpt corresponds to about double the basic frequency) are shown as pronounced minimums and maximums in the low-pass part of Figure 4C. The formants represent resonant frequencies in the vocal tract of the speaker. The clear presentability of the formants enables an adaptation and/or a selection of suitable sound components in the concatenative speech synthesis.

The huskiness of a voice can be determined in the low-pass part of the penultimate transformation stage (given 256 frequency values in the original signal:

TP7). The descent of the course of the curve between maximum Mx and minimum Mi characterizes the degree of the huskiness.

Said three speaker-specific characteristics are thus identified and can be intentionally influenced for the speech synthesis. It is thereby of particular significance that, in inverse wavelet transformation, the manipulation of a single speaker-specific characteristic influences only this; the other perceptibly relevant quantities remain unaffected. The basic frequency can thus be designationally adjusted without the huskiness of the voice being thereby influenced.

Another possible utilization is comprised in the selection of a suitable sound segment for concatenative linking with another sound segment, whereby the two sound segments were additionally recorded by different speakers in different contexts. With determination of spectral speech characteristics, a suitable sound segment to be linked can·be found since, with the characteristics, criteria are known that automatically enable a comparison of sound segments to one another according to specific rules and, thus, a selection of the suitable sound segment.

Figure 5 shows steps of a concatenative speech synthesis. A data bank is produced with a predetermined set of naturally spoken language of different speakers, whereby sound segments in the naturally spoken language are identified and stored. Numerous representatives of the various sound segments of a language derive that can be accessed by the data bank. The sound segments are, in particular, phonemes of a language or a concatenation of such phonemes. The possibilities in the compilation of new words are all the greater the smaller the sound segment is. Thus, the German language comprises a predetermined set of approximately 40 phonemes that suffice for the synthesis of nearly all words of the language. Different acoustic contexts are thereby to be taken into consideration dependent on the word in which the phoneme occurs. It is then important to embed the individual phonemes into the acoustic context such that discontinuities, which human hearing senses as unnatural and "synthetic", are avoided. As mentioned, the sound segments stem from different speakers and thus exhibit different speaker-specific characteristics. In order to synthesize an expression that has as natural an effect as possible, it is important to minimize the discontinuities. This can ensue be adaptation of the identifiable and modifiable speaker-specific characteristics or by selecting suitable sound segments from the data bank, whereby the speaker-specific characteristics likewise represent a critical aid in the selection.

By way of example, Figure 5 shows two sounds A 507 and B 508 that respectively exhibit individual sound segments 505 or, respectively, 506. The sounds A 507 and B 508 respectively derive from a spoken expression, whereby the sound A 507 is clearly distinct from the sound B 508. A parting line 509 indicates whereby the sound A 507 is to be linked to the sound B 508. In the present case, the first three sound segments of the sound A 507 are to be concatenatively linked with the last three sound segments of the sound B 508.

A temporal stretching or compression (see arrow 503) of the sound segments is implemented along the parting line 509 in order to avoid the discontinuous impression at the transition 509.

One version is comprised in an abrupt transition of the sounds parted along the parting line 509. However, said discontinuities that human hearing perceives as disturbing thereby occur. When, in contrast, a sound C is compiled [...] that the sound segments within a transition region 501 or 502 are considered, whereby a spectral distance criterion is adapted between two sound segments that can be allocated to one another in the respective transition region 501 or 502 (gradual transition between the sound segments). The Euclidean distance between the coefficients that are relevant in this region is utilized as the distance criterion, especially in the wavelet space.

Bibliography

[1] I. Daubechies, "Ten Lectures on Wavelets", Siam Verlag, 1992, ISBN 0-89871-274-2, Chapter 5.1, pages 129-137.